IST736 Text Mining

Professor Bei Yu

Fall 2021

HW\_2 Exploring Common Topics in Health Research News

The Health Research News Headlines (HW2\_headlines\_10000.csv) is a data set that sampled 10,000 headlines from health research press releases posted on the EurekAlert! website. You are going to explore the common themes of health research news by applying document clustering and topic modeling algorithms to this data set. This data set also contains timestamp for each headline such that you can analyze topic changes over the years using LDA.

The data set is already downloaded from the SU Experts database [1], cleaned and provided to you.

Deliverables:

1. Python script in .ipynb file
2. A word document in academic writing format to explain your methods and results, including
   1. Calculate cosine similarity between the third headline in the data set and the rest headlines, output the 10 most similar headlines to the first headline. Compare whether TF or TFIDF vectorization gives better result. You can choose other vectorization options (e.g. ngram, min or max TF or DF, etc.), and explain your choices.
   2. Use your best choice of vectorization options for K-Means clustering – explain how you find the best number of clusters, and what each cluster means.
   3. LDA topic modeling – explain how you find the best number of topics. When reporting the result, show the top 10 keywords for each topic, and give each topic a meaningful label if you can. Sometimes some topics can be hard to interpret. You can label those as “unsure”.
   4. [Optional 2 Extra points] Analyze the topic changes over the years using LDA.
   5. Note that kMeans and LDA are different algorithms. Do not tune one algorithm based on another algorithm’s result since there is no gold standard result to this problem.
3. Keep your report concise, i.e. no more than 4 pages, at least 12pt font and 1 inch margin all sides. Avoid large screenshots unless absolutely necessary.

Grading criteria:

1. Accuracy in method use.
2. Reasonable explanation of the results.